Community Dectection Assignment

David Terner

02/27/2022

I decided to use R and Rmarkdown to complete this assignment because I wanted to experiment with systematically changing the resolution parameter (for the Louvain community detection method). By contrast, Gephi makes it difficult to complete the same experiment. I use R's `ForceAtlas2` to achieve a Forced Altas layout and I use R's `resolution` to iteratively tune the resolution parameter.

Feel free to check out my code on Github.

Which resolution to use?

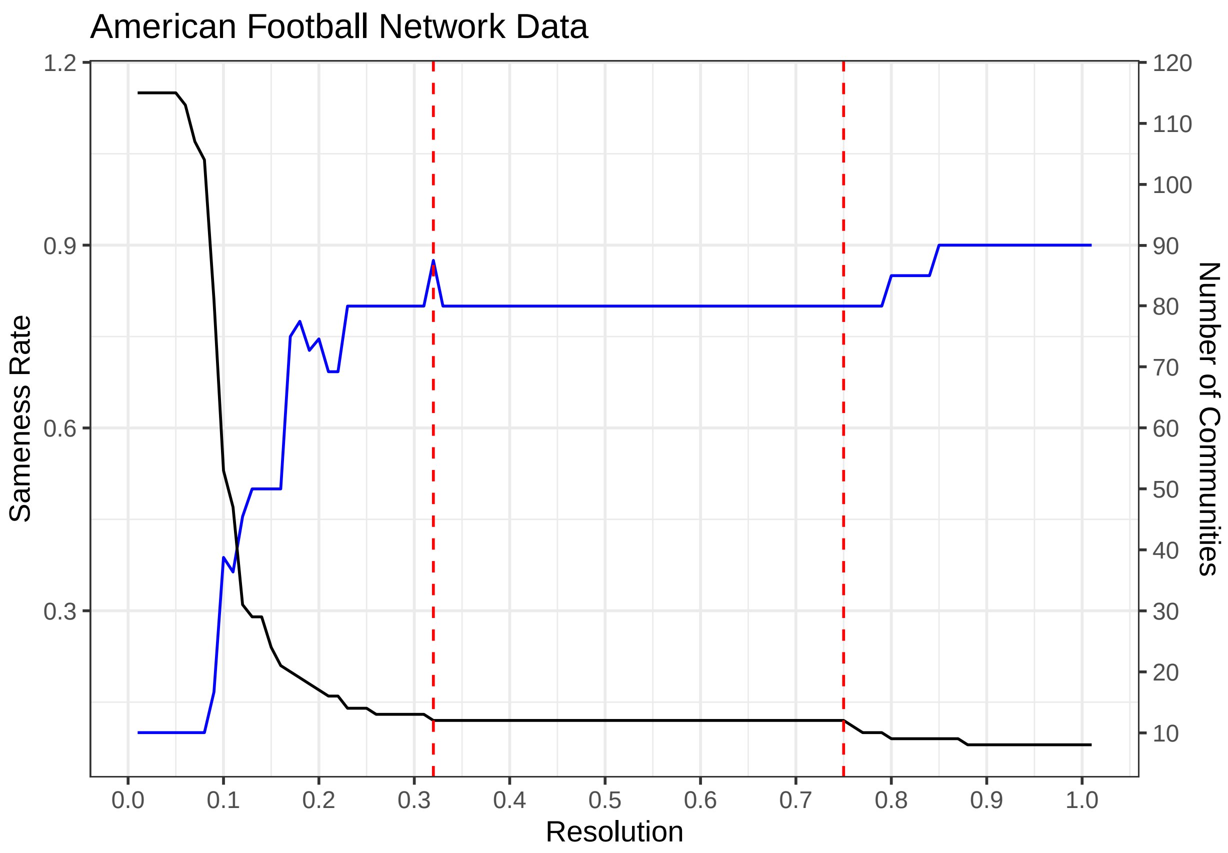
According to the `football.txt` file, college team/node values correspond to team conferences.

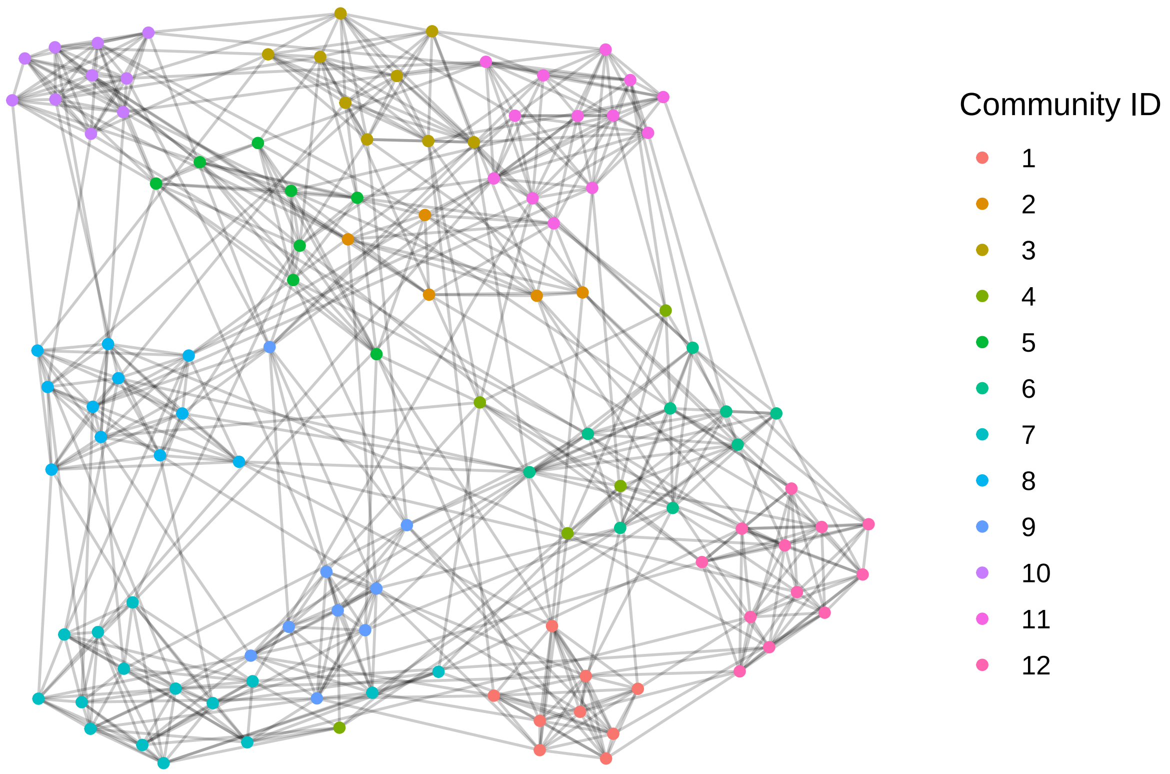
value	conference
0	Atlantic Coast
1	Big East
2	Big Ten
3	Big Twelve
4	Conference USA
5	Independents
6	Mid-American
7	Mountain West
8	Pacific Ten
9	Southeastern
10	Sun Belt
11	Western Athletic

I pick a resolution value t^* such that: (i) the number of communities equals the number of conferences, namely 12; and, (ii) each conference has, on average, the majority of its teams belong to to the same community. I operationalize this second requirement by first computing the share of top community (largest number of occurrences) per conference group, second determining the median share, and finally adjusting the resolution parameter such that this median share is maximized. Denote this median share value as the *sameness rate*.

I search over a grid of 100 values of the resolution parameter from 0.01 to 1.01 by 0.01 increments.

The figure below presents the *sameness rate*- measured by the blue line and left vertical axis- and the number of communities-measured by the black line and right vertical axis-in the same plot. The red dashed lines correspond to the smallest/largest resolution values that yield 12 communities. Based on the small blue spike that is intersected by the red dashed boundary line, I choose $t^* = 0.32$ as the resolution value.

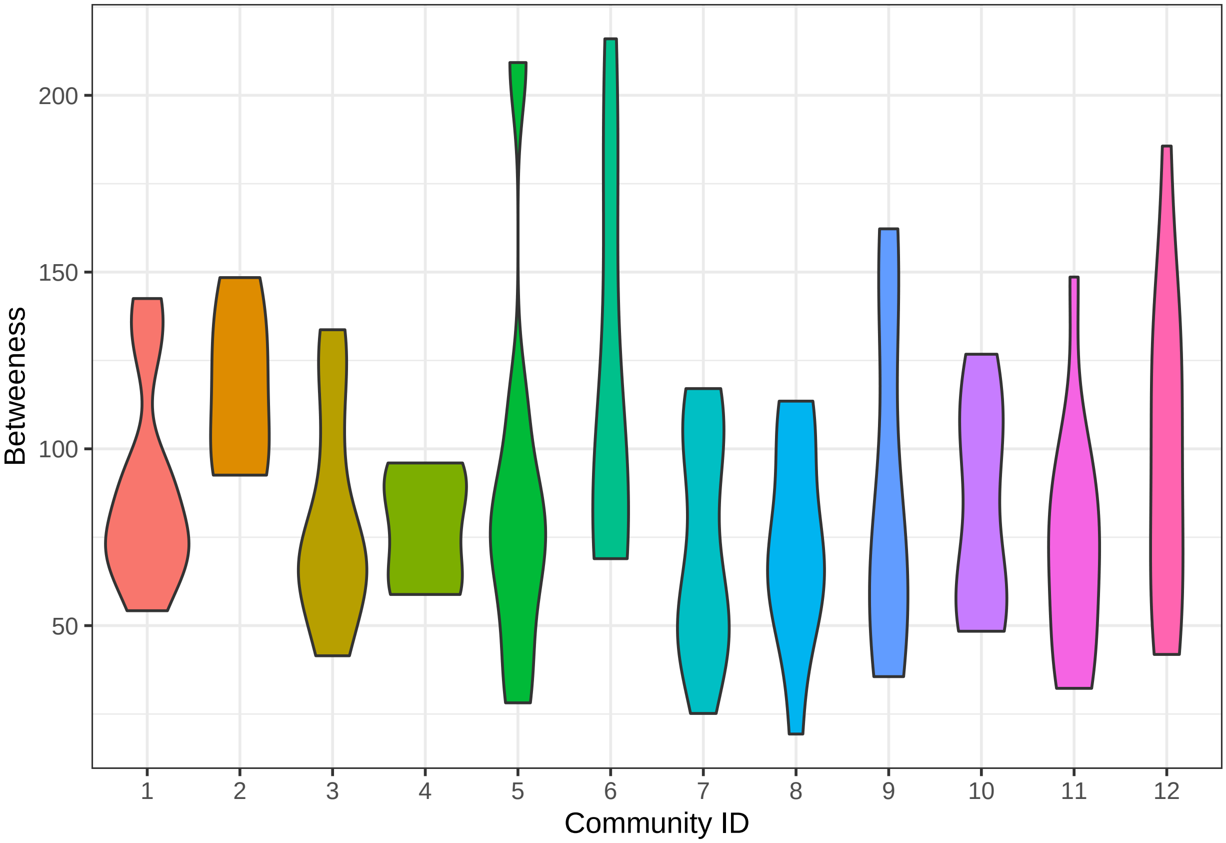

American Football Network Data
Sameness Rate
1.2
0.9
0.6
0.3
Number of Communities
120
110
100
90
80
70
60
50
40
30
20
10
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Resolution



Conference	Top Community Share
Atlantic Coast	1.00
Big Ten	1.00
Big Twelve	1.00
Mid-American	1.00
Mountain West	1.00
Pacific Ten	1.00
Southeastern	1.00
Conference USA	0.90
Big East	0.88
Western Athletic	0.80
Sun Belt	0.43
Independents	0.40

Using the Forced Atlas layout, the college football league does appear to have some community structure to it. Using the $t^* = 0.32$ resolution does a pretty good job matching communities to conferences: 7 out of 12 conferences consist of exactly one community. On the lower end, there is less of tight fit especially with the "Independents" conference. This shouldn't come as a shock given which schools belong to the "Independents" conference; Central Florida, Connecticut, Navy, Notre Dame, and Utah State are clearly a diverse group of programs with little in common save football conference membership. By contrast, the "Big 10" conference is exclusive to the MidWest.

How are betweenness centrality and community structure related?



Based on the above figure, there doesn't appear to be a strong relationship between between "betweeness" and "modularity". Nodes within a community can all have betweeness scores that relatively close to one another; see communities 4+5. However for other communities, such as 6 or 12, there's substantial betweeness variation.

Real network?

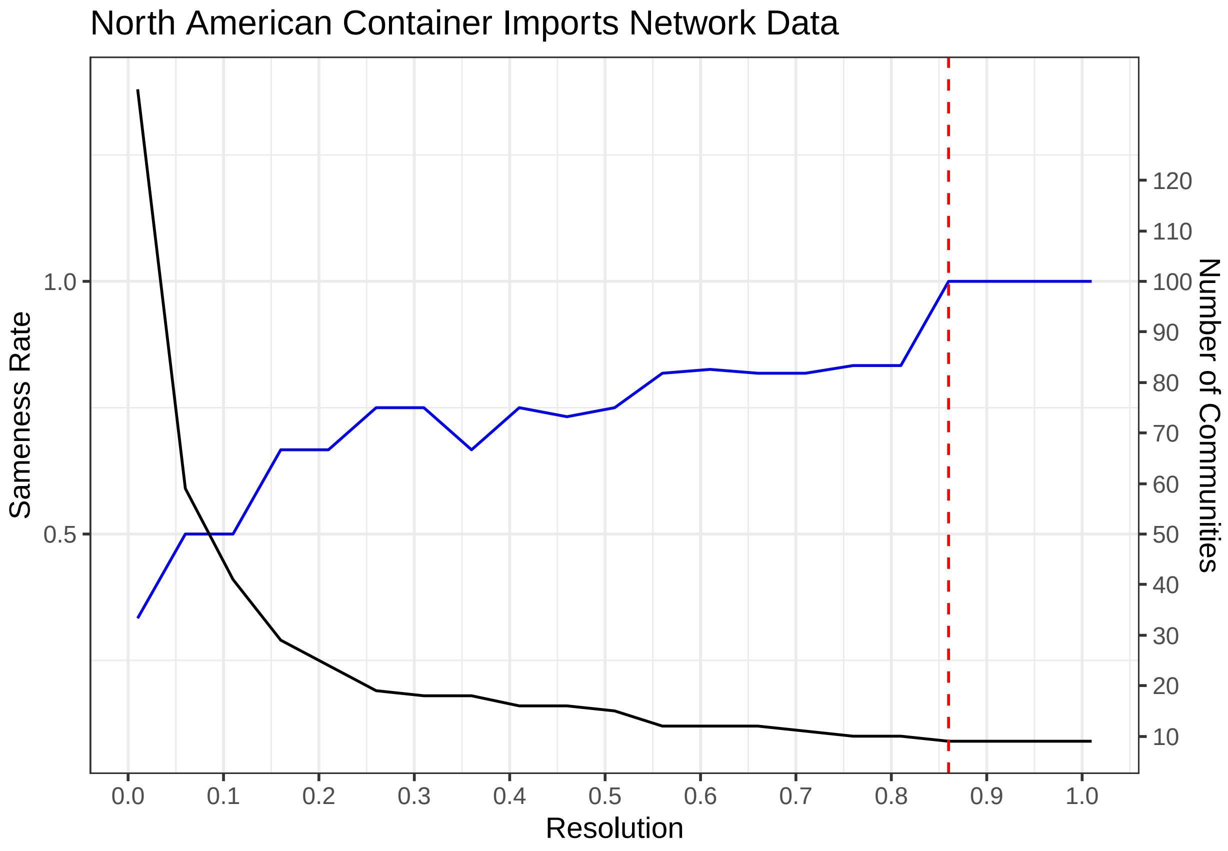
I use data my own data on containerized shipping schedules for 2015 North American imports from BlueWater Reporting. In this network, a node is a port and an edge represents a scheduled port-to-port connection. Edges are weighted by total traffic.

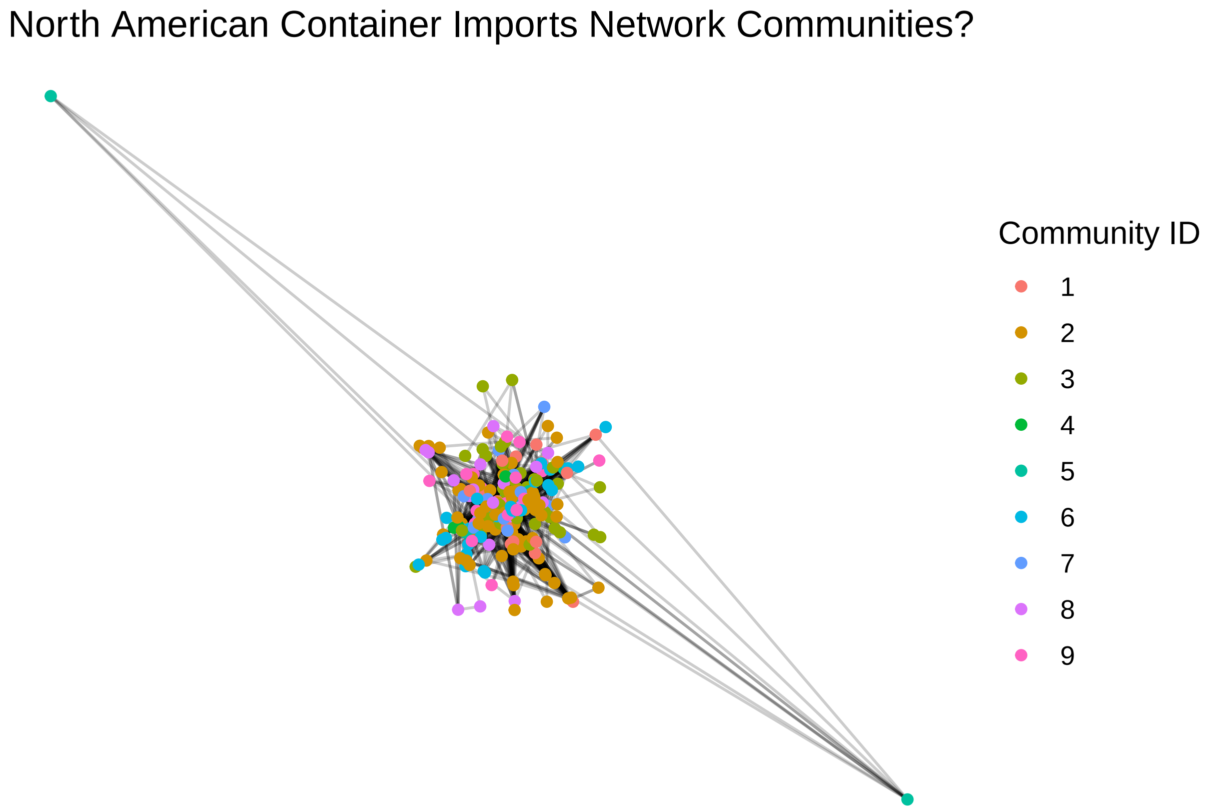
As with before, I iterate through a variety of different resolution parameters and choose $t* = 0.86$. At this resolution, I pick up nine (9) communities.

```
## Rows: 91692 Columns: 21

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (8): Origin, Destination, id, port_o, port_d, country_o, country_d, pai...
## dbl (13): time, trouble, lon_o, lat_o, lon_d, lat_d, length, panama, suez, y...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## [1] 116
```


North American Container Imports Network Data
Sameness Rate
1.0
0.5
Number of Communities
120
110
100
90
80
70
60
50
40
30
20
10
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Resolution

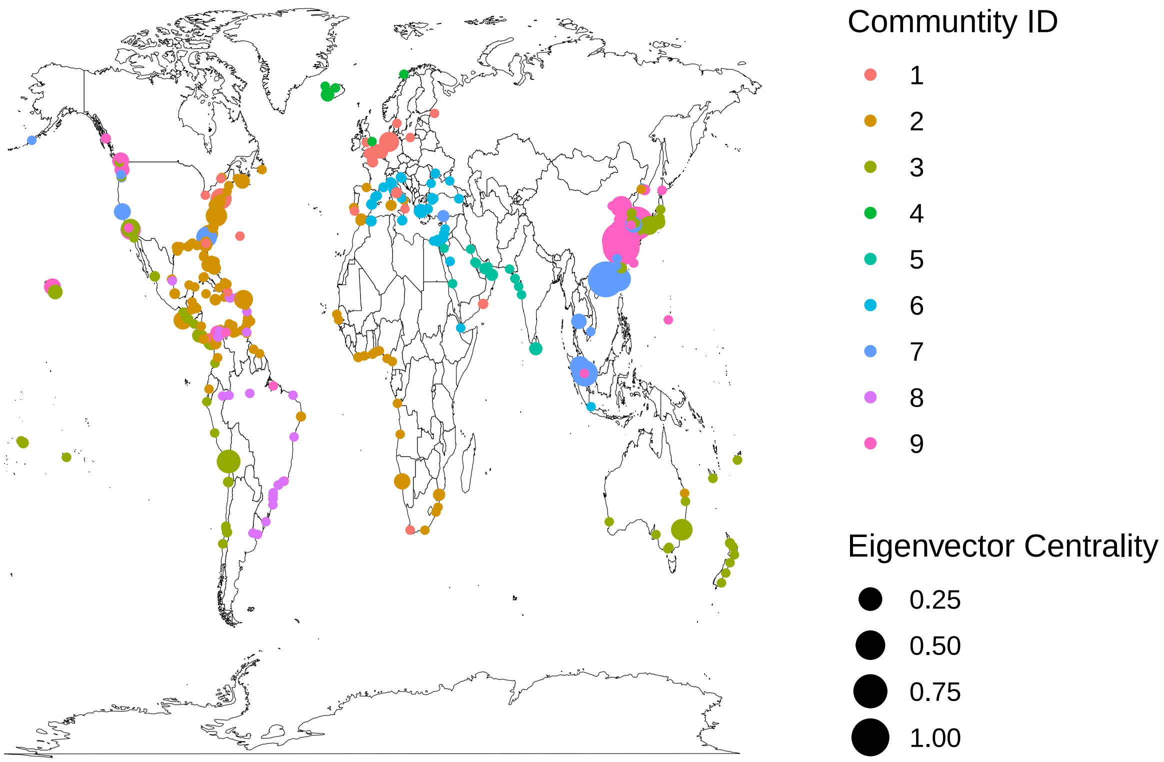


Without any additional detail, the preceding figure on port communities (e.g. position based on Forced Atlas, color based on membership) doesn't generate much intution abut the underlying network structure.

The final figure helps to crystalize the missing piece: geography. Community 1 roughly corresponds to Northen Chinese and South Korean ports + major West Coast ports. Community 2 represents Atlantic ocean facing ports. Community 3 is mostly "Ring of Fire" ports (e.g. the spine of the American Continent, Australia, and Japan). Community 4 is the North Atlantic. Community 5 is the Middle East and India. Community 6 corresponds with Mediterranean ports. Community 7 is South East Asia. Community 8 is Brazil and the Caribbean. Lastly, Community 9 doesn't have clear geographic orientation.

Ports are scaled in proportion to their eigenvector centrality. The size scale undescores the fact that South East Asia (Community 7) and Chinese and South Korean ports (Community 1) are home to the most centrally located ports. Put differently, port centrality and community stucture here tend to overlap pretty well, in part, becasue of the underyling geography of shipping schedules. Ports that are geographically proximate are likely to be on similar shippment rotations and are thus likely to be related to one another in a community sense.

Note that I excluded edges in this map to enhance the map's visual clarity.


Communtity ID
1
2
3
4
5
6
7
8
9
Eigenvector Centrality
0.25
0.50
0.75
1.00